03/12/2017

Investigate opinionated review document embeddings

Description:

Sentiment analysis is the computational study of people's appraisals and emotions toward entities, events and their attributes. The goal is to determine whether an opinionated document (e.g. reviews) or sentence expresses a positive or negative opinion. Opinions are important because whenever someone wants to make a decision, he would like to hear other's opinions [1]. The first step is generally to represent the text in a smaller dimensional space and then learn automatically hidden features from this representation. Finally, all these features are taken into account to compute the final sentiment polarity of the document.

However, it could be interesting to encode an opinionated document (e.g. review) in a smaller representation and reconstruct it given the latter. Autoencoders [2], sequence-to-sequence [3] or encoder-decoder [4] models might be used to achieve this goal. As a second step, we could try to find clusters of documents having the same sentiment polarity, generate new documents by using known points in the hidden space and observe how they differ from these points in term of text. Moreover, we might also observe how a document might change using a transformation (e.g. linear interpolation) between two points in this space (e.g. from positive to positive or negative to positive)?

The idea of this project would be to use a method to compute embeddings of short opinionated texts (e.g. restaurant reviews from *Yelp*[1], movie reviews from IMDb[2] or product reviews from Amazon[3]) by using encoder-decoder-like models. Secondly, investigate these embeddings and try to see if we can find clusters of same sentiment polarities, generate similar documents (by using only the decoder), etc. Finally, we could try to use this method as a data augmentation technique and thus, generate new training data and observe if there would be an improvement for a task of sentiment analysis.

Prerequisites:

- Knowledge about Machine Learning, especially neural networks (at least feedforward ones). You should have taken at least a ML course.
- Experienced with Python.
- (Optional) experienced with *Tensorflow* or *PyTorch*.

Supervisor: Diego Antognini (diego.antognini@epfl.ch)

References:

[1] https://www.cs.uic.edu/~liub/FBS/IEEE-Intell-Sentiment-Analysis.pdf
[2] https://arxiv.org/pdf/1506.01057.pdf
[3] https://arxiv.org/pdf/1705.03122.pdf
[4] http://aclweb.org/anthology/W17-1002

[1] https://www.yelp.com/dataset_challenge/dataset
[2] http://ai.stanford.edu/~amaas/data/sentiment/
[3] https://snap.stanford.edu/data/web-Amazon.html